P-2213

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-272686
(43)Date of publication of application : 08.10.1999

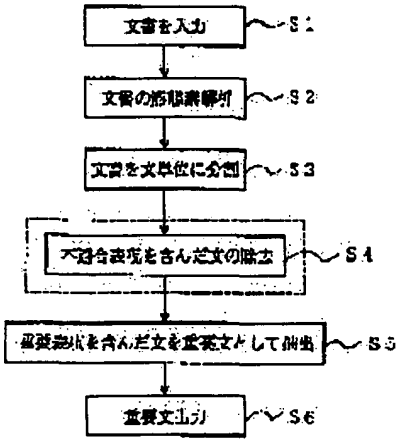
(51)Int.Cl. G06F 17/30
G06F 17/27

(21)Application number : 10-070688
(22)Date of filing : 19.03.1998

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>
(72)Inventor : HORII MUNEYUKI
MATSUOKA KOJI
TAKAGI SHINICHIRO

(54) METHOD AND DEVICE FOR IMPORTANT DOCUMENT SENTENCE EXTRACTION AND RECORD MEDIUM WHERE IMPORTANT DOCUMENT SENTENCE EXTRACTING PROGRAM IS RECORDED

(57)Abstract:
PROBLEM TO BE SOLVED: To easily extract an important sentence from a document with high precision.



SOLUTION: An improper expression table wherein improper expressions are described and an important expression table wherein important expressions are described are prepared; and a morpheme analysis (S2) of an inputted document is taken, the analyzed document is divided (S3) into sentences, and sentences including improper expressions are removed (S4) from the document divided into the sentences by referring to the improper expression table. From the document from which the sentences including the improper expression have been removed, sentences including important sentences are extracted (S5) as important sentence by referring to the important expression table. Here, the process for removing the sentences including the improper expressions is omitted in some cases and the sentences including the important expression may be extracted as important sentence directly from the document divided into the sentences by referring to the important expression table.

LEGAL STATUS

[Date of request for examination] 10.05.2001

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（A）　(11)特許出願公開番号
特開平11－272686
(43)公開日　平成11年(1999)10月8日

(51)Int.Cl.[6]	識別記号	FI		
G06F	17/30	G06F	15/401	320A
	17/27		15/20	550E
			15/38	D
			15/40	370A

審査請求　未請求　請求項の数5　OL　（全 7 頁）

(21)出願番号　特願平10－70688

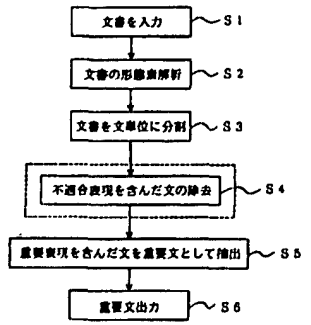
(22)出願日　平成10年(1998)3月19日

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目3番1号
(72)発明者　堀井　統之
東京都新宿区西新宿三丁目19番2号　日本電信電話株式会社内
(72)発明者　松岡　浩司
東京都新宿区西新宿三丁目19番2号　日本電信電話株式会社内
(72)発明者　高木　伸一郎
東京都新宿区西新宿三丁目19番2号　日本電信電話株式会社内
(74)代理人　弁理士　鈴木　誠

(54)【発明の名称】　文書重要文抽出方法、文書重要文抽出装置及び文書重要文抽出プログラムを記録した記録媒体

(57)【要約】
【課題】　文書中の重要文を、簡単に且つ高い精度で抽出することを可能にする。
【解決手段】　重要文として不適な表現を記述した不適表現テーブルと、重要表現を記述した重要表現テーブルを用意し、入力された文書を形態素解析し（Ｓ２）、該形態素解析された文書を文単位に分割し（Ｓ３）、該文単位に分割された文書中から不適表現テーブルを参照して不適表現を含んだ文を取り除き（Ｓ４）、該不適表現を含んだ文を取り除いた後の文書中から重要表現テーブルを参照して重要表現を含んだ文を重要文として抽出する（Ｓ５）。ここで、場合によっては、不適表現を含んだ文を取り除く処理は省略し、文単位に分割された文書中から、重要表現テーブルを参照して直接重要表現を含んだ文を重要文として抽出することでもよい。

本発明の文書重要文抽出方法の処理フロー図



【特許請求の範囲】
【請求項１】　文書を入力し、該入力された文書を形態素解析し、該形態素解析された文書を文単位に分割し、重要表現を記述した重要表現テーブルを参照して、該文単位に分割された文書中から重要表現を含んだ文を重要文として抽出することを特徴とする文書重要文抽出方法。
【請求項２】　請求項１記載の文書重要文抽出方法において、重要文として不適な表現を記述した不適表現テーブルを参照して、文書中から不適表現を含んだ文を取り除いた後の文書を重要文抽出の対象とすることを特徴とする文書重要文抽出方法。
【請求項３】　文書を入力する手段と、前記入力された文書を形態素解析する手段と、前記形態素解析された文書を文単位に分割する手段と、重要表現を記述した重要表現テーブルと、前記重要表現テーブルを参照して、前記文単位に分割された文書中から重要表現を含んだ文を重要文として抽出する手段を有することを特徴とする文書重要文抽出装置。
【請求項４】　文書を入力する手段と、前記入力された文　を形態素解析する手段と、前記形態素解析された文書を文単位に分割する手段と、重要文として不適な表現を記述した不適表現テーブルと、前記不適表現テーブルを　照して、前記文単位に分割された文書中から不適表現を含んだ文を取り除く手段と、重要表現を記述した重要表現テーブルと、前記重要表現テーブルを参照して、前記不適表現を含んだ文を取り除いた文書中から重要表現を含んだ文を重要文として抽出する手段を有することを特徴とする文書重要文抽出装置。
【請求項５】　文書から重要文を抽出するための文書重要文抽出プログラムを記録したコンピュータ読み取り可能な記録媒体であって、入力された文書を形態素解析する処理プロセスと、形態素解析された文書を文単位に分割する処理プロセスと、重要文として不適な表現を記述した不適表現テーブルを参照して、文単位に分割された文書中から不適表現を含んだ文を取り除く処理プロセスと、重要表現を記述した重要表現テーブルを参照して、不適表現を含んだ文を取り除いた文書中から重要表現を含んだ文を重要文として抽出する処理プロセスを有することを特徴とする記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、コンピュータによる文書処理に係り、詳しくは、文書の要約等を目的として文書の重要文を抽出する文書重要文抽出方法、文書重要文抽出装置、及びそのための文書重要文抽出プログラムを記録した記録媒体に関する。
【０００２】
【従来の技術】大量の情報、特にWWW（World Wide Web）の閲覧情報や電子メールといった電子的な情報を効率良く概観し、必要な情報だけを取得するためには、文書の要約技術が必要不可欠である。
【０００３】従来、コンピュータによる文書処理において、文書を要約するために文書中から重要文を自動を抽出する技術としては、「文書中の単語頻度を利用する手法」、「文書構造を解析する手法」がよく用いられる。「文書中の単語頻度を利用する手法」は、大量の文書を統計処理することにより、文書中に現れる各単語を単語単位の出現頻度、文書単位の出現頻度を基にランキングし、そのランキングが高い単語を含む文を重要文として選択する手法である。一方、「文書構造を解析する手法」は、文書の構造から主題があるべきブロックを同定し、その中から重要文を選択したり、話題の導入、転換等を表す言語表現を指標に重要文を選択する手法である。
【０００４】
【発明が解決しようとする課題】従来技術の「文書中の単語頻度を利用する手法」は、大量の文書データが用意できる場合には有効な方法となるが、反面、（１）大量の文書データが必要である、（２）必ずしもランキングが高い単語（一般的に文書内出現頻度が高く、他文書における出現頻度が少ない単語）が重要文に含まれているとは限らない、（３）該当単語の出現頻度が非常に高い場合、ほとんどの文に該当単語が含まれて多数の文が選択されてしまう、などの問題点がある。
【０００５】また、「文書構造を解析する手法」は、文書構造をうまく解析できれば効果的であるが、文書構造を正しく解析できなければならず、話し言葉のように解析が困難な文書や、電子メールのように定型的な構造を持たない文書を対象とする場合、重要文抽出までたどりつかない問題がある。
【０００６】本発明は、上記従来技術の問題点に鑑みてなされたもので、文書中から簡単かつ高い精度で、重要表現を含んだ文を重要文として抽出する文書重要文抽出方法、文書重要文抽出装置、及びそのための文書重要文抽出プログラムを記録した記録媒体を提供することを目的とする。
【０００７】
【課題を解決するための手段】上記目的を達成するために、本発明は、重要文として不適な表現と重要表現を記述したテーブルを用意し、文書中から不適表現を含んだ文を取り除いた後の文書を対象に、重要表現を含んだ文を重要文として抽出することを特徴とするものである。なお、場合によっては、不適表現を含んだ文を取り除く処理は省略してもよい。
【０００８】
【発明の実施の形態】図１は、本発明の文　重要文抽出方法の処理フロー図である。図１に示すように、本発明方法では、文書を入力し（ステップＳ１）、該入力された文書を形態素解析し（ステップＳ２）、該形態素解析

された文書を文単位に分割し（ステップＳ３）、まず、重要文として不適な表現を記述した不適表現テーブルを参照して、文単位に分割された文書中から不適表現を含んだ文を取り除き（ステップＳ４）、次に、重要表現を記述した重要表現テーブルを参照して、不適表現を含んだ文を取り除いた文書中から重要表現を含んだ文を重要文として抽出し（ステップＳ５）、該抽出した重要文を出力する（ステップＳ６）。ここで、ステップＳ４の不適表現を含んだ文を取り除く処理は、場合によっては省略し、文単位に分割された文書中から、重要表現テーブルを参照して直接重要表現を含んだ文を重要文として抽出してもよい。
【０００９】図２は、本発明の文書重要文抽出装置の一実施例の概略ブロック図である。本文書重要文抽出装置は、文書入力装置１０、形態素解析部２１、単文分割部２２、不適文除去部２３及び重要文抽出部２４の各処理機能を有する文書処理装置本体２０、文書出力装置３０、及び、単語辞書４１、不適表現テーブル４２、及び重要表現テーブル４３を格納する記憶装置４０などで構成される。この構成は、所謂コンピュータシステムで実現されるものである。ここで、文書入力装置１０は、キーボード、イメージスキャナ、ＦＤーやＣＤ－ＲＯＭ等のドライバ、あるいは、WWWや電子メールでの文書入力のための通信インタフェースの総称である。文書処理装置本体２０は、実行するプログラムや所要データ等を格納するための内部メモリを有する所謂ＣＰＵ本体である。文書出力装置３０はディスプレイやプリンタである。記憶装置４０はハードディスクなどである。
【００１０】図２において、文書入力装置１０は、重要文抽出の対象となる文書の入力を行い、入力された文書を形態素解析部２１に送出する。形態素解析部２１は、文書入力装置１０より受け取った文書を単語辞書４１を参照して形態素解析を行い、この形態素解析された文書を単文分割部２２に送出する。単文分割部２２は、形態素解析部２２より受け取った文書を文単位に分割する。ここまでは従来と同様である。単文分割部２２では、この文単位に分割された文書を不適文除去部４０に送出する。
【００１１】不適文除去部２３は、不適表現テーブル４２を参照することにより、単文分割部２２より受け取った文書中から不適表現を含んだ文を取り除く。不適表現テーブル４２には、重要文として不適な表現が形態素情報、文位置情報として記述されている。不適文除去部２３により、不適表現テーブル４２に記述されたいずれかの表現を含んだ文全てが文書中から取り除かれる。不適表現を含んだ文を取り除いた文書が重要文抽出部２４に送出される。
【００１２】重要文抽出部２４は、重要表現テーブル４３を参照することにより、不適文除去部２３より受け取った文書中から重要表現を含んだ文を抽出する。重要表現テーブル４３には、重要表現が形態素情報、文位置情報として記述されている。重要文抽出部２４により、重要表現テーブル４３に記述されたいずれかの表現を含んだ文全てが重要文として文書中から抽出される。抽出された重要文は文書出力装置３０に出力される。
【００１３】次に、具体例を用いて、本発明による文書から重要文を抽出する手法を説明する。
【００１４】文書入力装置１０で、以下の電子メール文書が入力された場合を考える。
＜入力＞
横須賀情報通信　高木様
三田商事の竹石です。いつもお世話になっております。
システム緊急会議のお願いです。
貴社に納入して頂いた在庫管理システムが今朝ダウンし、現在非常モードで稼働中です。主な状況は以下の通りです。
・端末からアクセスしても初期画面が表示されません。
・バックアップデータがリストアできません。
年末時期で業務に支障があり、早急に対応策を検討しなければなりません。至急電話を頂けないでしょうか？本日１５時から１７時に三田商　本社会議室で緊急対策会議を行いますので、ご出席をお願いしたいと思います。
申し訳ありませんが、よろしくお願いします。
【００１５】形態素解析部２１では、単語辞書４１を照して上記入力文書を形態素解析する。たとえば、上記入力文書のうち、「三田商事の竹石です。」を形態素解析した結果は、図３のようになる。図３において、上段が原文及び分割点、中段が各単語の標準表記、下段が各単語の品詞を表す。文書入力装置１０で入力された文書は全て形態素解析部２１により、図３と同様に形態素解析される。
【００１６】単文分割部２２では、形態素解析された入力文書を文単位に分割する。上記入力文　は、以下のように文単位に分割される。ただし、以下では、説明の都合上、分割された文の順番に文番　を付与して示す。また、形態素情報は省略する。
１　横須賀情報通信　高木様
２　三田商事の竹石です。
３　いつもお世話になっております。
４　システム緊急会議のお願いです。
５　貴社に納入して頂いた在庫管理システムが今朝ダウンし、現在非常モードで稼働中です。
６　主な状況は以下の通りです。
７　・端末からアクセスしても初期画面が表示されません。
８　・バックアップデータがリストアできません。
９　年末時期で業務に支障があり、早急に対応策を検討しなければなりません。
１０　至急電話を頂けないでしょうか？
１１　本日１５時から１７時に三田商事本社会議室で緊

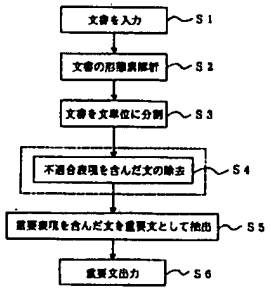
急対策会議を行いますので、ご出席をお願いしたいと思います。
１２　申し訳ありませんが、よろしくお願いします。
【００１７】不適文除去部２３では、単文分割部２２で分割された全ての文に対し、不適表現テーブル４２に記述された表現のいずれかが含まれているかどうかを調べ、１つでも含まれていた場合、その文を、不適表現を含んだ文、すなわち重要文とはなり得ない文として、単文分割部２２から受け取った文書から取り除く。
【００１８】図４に、不適表現テーブル４２の例を示す。不適表現テーブル４２は、文位置情報と形態素情報から構成される。文位置情報は、形態素情報で記述されている表現が含まれているかどうかをチェックする対象となる文位置を記述している。通常"＊"となっており、全ての文を対象とする。たとえば"先頭文"となっていれば、先頭の文しか対象とならない。形態素情報は１組以上の接続情報、表層表記、標準表記、品詞の４つ組から構成される。接続情報は、次の形態素の位置を規定する。"ｎｅｘｔ"は次の形態素が直後にあることを表し、"ｆａｒ"は次の形態素が離れていても良いことを表す。また、"ｅｎｄ"は文末であることを表し、"－"は次に形態素情報がないことを表す。表層表記、標準表記、品詞における"＊"は全てにマッチすることを表す。
【００１９】上記分割された各文において、たとえば、
１　横須賀情報通信　高木様
では、「様（文末）」の部分が不適表現テーブル４２のテーブルＩＤ１００の表現にマッチする。従って、文書から取り除かれる文となる。同様にして、
２　三田商事の竹石です。
は、「竹石です」の部分がテーブルＩＤ１０１にマッチ、
３　いつもお世話になっております。
は、「世話になっ」の部分がテーブルＩＤ１０２にマッチ、
１２　申し訳ありませんが、よろしくお願いします。
は、「が、よろしくお願いし」の部分がテーブルＩＤ１０３にマッチする。なお、文番号１２は、「、」と「お願い」の間に「よろしく」が入っているが、テーブルＩＤ１０３の「、」の接続情報が"ｆａｒ"になっているので、「お願い」が離れていても構わない。
【００２０】このようにして、不適文除去部２３では、入力文書中から文番号１、２、３、１２の４文が取り除かれ、以下の文書が重要文抽出部２３へ送出される。
４　システム緊急会議のお願いです。
５　貴社に納入して頂いた在庫管理システムが今朝ダウンし、現在非常モードで稼働中です。
６　主な状況は以下の通りです。
７　・端末からアクセスしても初期画面が表示されません。
８　・バックアップデータがリストアできません。
９　年末時期で業務に支障があり、早急に対応策を検討しなければなりません。
１０　至急電話を頂けないでしょうか？
１１　本日１５時から１７時に三田商事本社会議室で緊急対策会議を行いますので、ご出席をお願いしたいと思います。
【００２１】重要文抽出部２４では、不適文除去部２３４０から受け取った、不適表現を含んだ文の取り除かれた文書中の全ての文に対し、重要表現テーブル４３に記述された表現のいずれかが含まれているかどうかを調べ、１つでも含まれていた場合、重要文として抽出する。
【００２２】図５に、重要表現テーブル４３の例を図４に示す。重要表現テーブル４３の構成は、図４の不適表現テーブル４２の構成と同じである。したがって、重要文抽出部２４では不適文除去部２３と同様の手法で照合を行う。
【００２３】上記文例では、不適文除去部２３から受け取った文書のうち、
４　システム緊急会議のお願いです。
の「お願いです」の部分がテーブルＩＤ２０２にマッチ、
１０　至急電話を頂けないでしょうか？
の「か？」の部分がテーブルＩＤ２００にマッチ、
１１　本日１５時から１７時に三田商事本社会議室で緊急対策会議を行いますので、ご出席をお願いしたいと思います。
の「お願いし」の部分がテーブルＩＤ２０１にマッチする。なお、テーブルＩＤ２０４の形態素情報は、
９　年末時期で業務に支障があり、早急に対応策を検討しなければなりません。
の「を」の部分とマッチするが、文位置情報に"先頭文"とあるので、先頭文でない文番号９は該当しない。（この場合の先頭文は文番号４）。
【００２４】以上により、重要文抽出部２４では、上記の３文、すなわち、
４　システム緊急会議のお願いです。
１０　至急電話を頂けないでしょうか？
１１　本日１５時から１７時に三田商事本社会議室で緊急対策会議を行いますので、ご出席をお願いしたいと思います。
が重要文として抽出されることになる。
【００２５】図２の実施例では、不適文除去部２３で不適表現を含んだ文を取り除いた文書を対象として、重要文抽出部２４で重要文を抽出しているが、不適文除去部２３での処理を省略し、不適文を取り除かずに重要文抽出を行うこともできる。その場合、上記文　例では、文番号１から１２の文全てが重要文抽出部５０への入力となるので、文番号４、１０、１１の３文以外に、

１２　申し訳ありませんが、よろしくお願いします。
が抽出される。（「お願いし」とテーブルＩＤ２０１がマッチ）。
【００２６】したがって、抽出される重要文は、
４　システム緊急会議のお願いです。
１０　至急電話を頂けないでしょうか？
１１　本日１５時から１７時に三田商事本社会議室で緊急対策会議を行いますので、ご出席をお願いしたいと思います。
１２　申し訳ありませんが、よろしくお願いします。の４文となる。
【００２７】このように、不適文除去部２３の処理を省略することで、それほど重要ではない文番号１２も抽出されてしまうが、上記文書例では、多少精度が落ちる程度で問題はない。
【００２８】次に、他の具体例として、文書入力装置１０で、以下の電子メール文書が入力された場合を考える。
＜入力＞
田中です。
４月２５日に「第３回情報通信学習会」を開催します。
今回のテーマは、「エージェント通信」です。佐藤さんのグループには関連の深いテーマだと思います。参加希望者がございましたら、田中までご連絡ください。多数の参加、お待ちしております。
【００２９】形態素解析部２１、単文分割部２２を経て、
１　田中です。
２　４月２５日に「第３回情報通信学習会」を開催します。
３　今回のテーマは、「エージェント通信」です。
４　佐藤さんのグループには関連の深いテーマだと思います。
５　参加希望者がございましたら、田中までご連絡ください。
６　多数の参加、お待ちしております。
が不適文除去部２３に送出される。
【００３０】不適文除去部２３では、図４の不適表現テーブル４２を参照することにより、
１　田中です。
が取り除かれる。（「田中です」とテーブルＩＤ１０１がマッチ）。
【００３１】したがって、重要文抽出部２４には、
２　４月２５日に「第３回情報通信学習会」を開催します。
３　今回のテーマは、「エージェント通信」です。
４　佐藤さんのグループには関連の深いテーマだと思います。
５　参加希望者がございましたら、田中までご連絡ください。
６　多数の参加、お待ちしております。が送出される。
【００３２】重要文抽出部２４では、図５の重要表現テーブル４３を参照することにより、
２　４月２５日に「第３回情報通信学習会」を開催します。
５　参加希望者がございましたら、田中までご連絡ください。
がそれぞれ「を」（先頭文）とテーブルＩＤ２０４のマッチ、「ご連絡ください」とテーブルＩＤ２０３のマッチで重要文として抽出される。
【００３３】この例で不適文除去部２３の処理を省略した場合、重要文抽出部２４では、文番号２が抽出されない。これは文番号２が先頭文でなくなるため、テーブルＩＤ２０４の文位置情報を満たさないからである。したがって抽出される重要文は、
５　参加希望者がございましたら、田中までご連絡ください。
となり、精度がかなり落ちることになる。
【００３４】以上、本発明の実施例について説明した。ここで、図１に示した本発明の文書重要文抽出方法の処理フローは、文書重要文抽出プログラムとして、ＦＤもしくはＣＤ－ＲＯＭ等の記録媒体にコンピュータ読み取り可能な形式で記録し、販売してもよい。この記録媒体に記録された文書重要文抽出プログラムを計算機にインストールすると、図２の実施形態での使用が可能になる。
【００３５】
【発明の効果】上述のように本発明によれば、重要文として不適な表現と重要表現を記述したテーブルを用意するだけで、簡単に、また高い精度で文書中の重要文を抽出することができる。本発明は、特に電子メール文書等に有効である。
【図面の簡単な説明】
【図１】本発明の文書重要文抽出方法の処理フロー図である。
【図２】本発明の文書重要文抽出装置の一実施例の概略ブロック図である。
【図３】形態素解析結果の具体例を示す図である。
【図４】本発明に用いる不適表現テーブルの内容例を示す図である。
【図５】本発明に用いる重要表現テーブルの内容例を示す図である。
【符号の説明】
１０　文書入力装置
２０　文書処理装置本体
２１　形態素解析部
２２　単文分割部
２３　不適文除去部
２４　重要文抽出部
３０　文書出力装置

４０　記憶装置
４１　単語辞
４２　不適表現テーブル
４３　重要表現テーブル

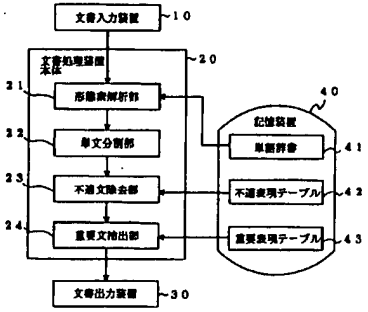
【図１】

本発明の文書重要文抽出方法の処理フロー図



【図２】

本発明の文書重要文抽出装置の一実施例の概略ブロック図



【図３】

形態素解析結果の具体例

三田	/商事	/の	/竹石	/です	/。/	
三田	商事	の	竹石	だ	。	
固有名詞(場所)	普通名詞	接続助詞	固有名詞(姓)	判定詞	句点	

【図４】

本発明に用いる不適表現テーブルの内容例

42

テーブルID	文位置情報	形態素情報			
		接続情報	表層表記	標準表記	品詞
100	*	end	終	終	名詞性名詞接尾
101	*	next	*	*	固有名詞（姓）
		–	*	だ	判定詞
102	*	next	住部	住部	普通名詞
		next	に	に	格助詞
		–	*	なる	動詞
103	*	next	が	が	接続助詞
		fer	、	、	読点
		next	お願い	お願い	普通名詞
		–	*	する	動詞

【図５】

本発明に用いる重要表現テーブルの内容例

43

テーブルID	文位置情報	形態素情報			
		接続情報	表層表記	標準表記	品詞
200	*	next	か	か	終助詞
		–	？	？	記号
201	*	next	お願い	お願い	普通名詞
		–	*	する	動詞
202	*	next	お願い	お願い	普通名詞
		–	*	だ	判定詞
203	*	next	*	*	普通名詞
		–	ください	くださる	動詞
204	先頭文	–	を	を	格助詞